Customer Segmentation

[1] Yamini Chouhan, [2] T Vaishnavi, [3] R Dinesh, [4] B Karthik
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
yaminichouhan_cse@siddhartha.co.in, 23tq1a5621@siddhartha.co.in, 23tq1a5646@siddhartha.co.in, 23tq1a5657@siddhartha.co.in,

Abstract

Customer segmentation is a vital strategy in modern e-commerce that enables businesses to better understand their customers, deliver personalized experiences, and optimize marketing strategies. This project presents a machine learning-based approach to customer segmentation using the K-Means clustering algorithm on a synthetic e-commerce dataset containing demographic, geographic, behavioral, and campaign interaction features.

The dataset consists of 10,000 customer records with 42 attributes, providing a comprehensive view of customer profiles. The methodology includes data preprocessing, exploratory data analysis (EDA), and feature engineering to prepare the data for clustering. The Elbow Method is used to determine the optimal number of clusters, resulting in the identification of four distinct customer segments.

Each segment is analyzed based on key characteristics such as age, annual income, total spending, family size, and purchasing behavior. The segmentation provides valuable insights that can be used for targeted marketing, personalized recommendations, and improving customer retention strategies.

The results demonstrate that unsupervised machine learning techniques like K-Means clustering are highly effective in identifying meaningful customer groups from complex datasets. This approach can further support advanced applications such as customer lifetime value analysis, churn prediction, and campaign optimization.

I. Introduction

Customer segmentation has become a crucial aspect of modern e-commerce due to the rapid growth of online consumers and the increasing diversity in customer behavior. Businesses today operate in highly competitive environments where understanding customer needs and preferences is essential for success. However, the heterogeneous nature of customer data makes it challenging for organizations to effectively analyze and categorize their customer base. Traditional mass marketing approaches are no longer sufficient, as they fail to address individual preferences, leading to reduced customer engagement, lower conversion rates, and decreased brand loyalty.

One of the major challenges faced by e-commerce businesses is the effective utilization of vast amounts of data generated through customer interactions. This data includes purchase history, browsing behavior, demographic information, geographic location, and responses to marketing campaigns. Without proper analytical techniques, such data remains underutilized, preventing businesses from gaining valuable insights.

As a result, organizations struggle to identify high-value customers, predict customer churn, personalize marketing strategies, and optimize resource allocation. They also face difficulties in designing targeted campaigns and understanding cross-selling and up-selling opportunities within different customer groups. These challenges ultimately impact business performance and profitability.

II. Literature Survey

Customer segmentation has evolved significantly over the years, transitioning from simple demographic-based grouping to advanced machine learning-driven approaches. The concept was first introduced by Smith (1956), who defined market segmentation as the process of dividing a market into homogeneous groups.

A major advancement in segmentation theory was introduced by Yankelovich (1964), who emphasized behavioral segmentation. This approach recognized that customers with similar demographic profiles could exhibit different purchasing patterns and preferences. This insight laid the foundation for modern multi-dimensional segmentation techniques that combine demographic, behavioral, and transactional data.

RFM Analysis in Customer Segmentation

RFM (Recency, Frequency, Monetary) analysis has been a widely adopted technique in customer segmentation. Hughes (1994) formalized the RFM framework, highlighting its effectiveness in evaluating customer value and engagement. Later, Fader, Hardie, and Lee (2005) enhanced this model by integrating probabilistic methods to estimate customer lifetime value.

Machine Learning Approaches to Customer Segmentation

- **K-Means Clustering:**
  K-Means is one of the most widely used algorithms due to its simplicity, efficiency, and scalability. Arthur and Vassilvitskii (2007) introduced K-Means++, which improves cluster initialization and enhances clustering performance. Studies by Kanungo et al. (2002) confirmed its effectiveness for handling large-scale customer datasets.
- **Hierarchical Clustering:**
  Hierarchical clustering methods, introduced by Johnson (1967), provide a tree-like structure of clusters, allowing deeper insights into customer group relationships. However, its high computational cost makes it less suitable for very large datasets.
- **DBSCAN (Density-Based Clustering):**
  DBSCAN, proposed by Ester et al. (1996), is effective in identifying clusters of arbitrary shapes and handling noise or outliers. Research by Birant and Kut (2007) demonstrated its usefulness in customer segmentation where irregular patterns exist.
- **Gaussian Mixture Models (GMM):**
  GMM is a probabilistic clustering approach that assigns data points to clusters based on probability distributions.

III. System Analysis

System analysis focuses on understanding the requirements and workflow of the customer segmentation system. The system processes customer data such as demographic details, geographic information, purchase history, and campaign interactions. Data preprocessing techniques like handling missing values, encoding categorical variables, and normalization are applied to prepare the dataset. Exploratory Data Analysis (EDA) is performed to identify patterns and relationships among features. Feature engineering is used to create meaningful attributes such as total spending and customer activity levels. The processed data is then used for clustering. The K-Means algorithm is applied to group customers into distinct segments. The optimal number of clusters is determined using methods like the Elbow Method. The system generates insights based on each cluster's characteristics. Overall, the system supports data-driven decision-making for targeted marketing strategies.

Existing System

The existing system for customer segmentation mainly relies on traditional methods such as demographic grouping and manual analysis. Businesses often segment customers based on basic attributes like age, gender, and location. These methods do not capture complex customer behaviors and preferences. Marketing strategies are usually designed using broad assumptions rather than data-driven insights. Some organizations use basic statistical tools, but they lack advanced analytical capabilities. The existing systems are often time-consuming and require manual effort. They do not efficiently handle large volumes of customer data. Additionally, they lack automation and real-time processing. As a result, segmentation is often inaccurate and less effective. This leads to poor targeting and reduced marketing efficiency.

Disadvantages of Existing System

- Limited to basic demographic segmentation
- Inability to capture complex customer behavior
- Lack of data-driven decision-making
- Time-consuming and manual analysis
- Inefficient handling of large datasets
- No real-time or dynamic segmentation

Proposed System

The proposed system uses machine learning techniques for advanced customer segmentation. It processes customer data including demographic, behavioral, and transactional features. Data preprocessing and feature engineering are applied to improve data quality and relevance. The K-Means clustering algorithm is used to group customers into meaningful segments. The optimal number of clusters is determined using the Elbow Method. Each cluster is analyzed based on key characteristics such as income, spending habits, and engagement levels. The system automatically identifies patterns in customer behavior. It provides actionable insights for targeted marketing and personalization. The system is scalable and can handle large datasets efficiently. Overall, it enables businesses to make informed and strategic decisions.

Advantages of Proposed System

- Accurate and data-driven customer segmentation
- Ability to handle large and complex datasets
- Identifies hidden patterns in customer behavior
- Supports personalized marketing strategies
- Improves customer engagement and retention
- Enhances targeting for campaigns
- Saves time through automation

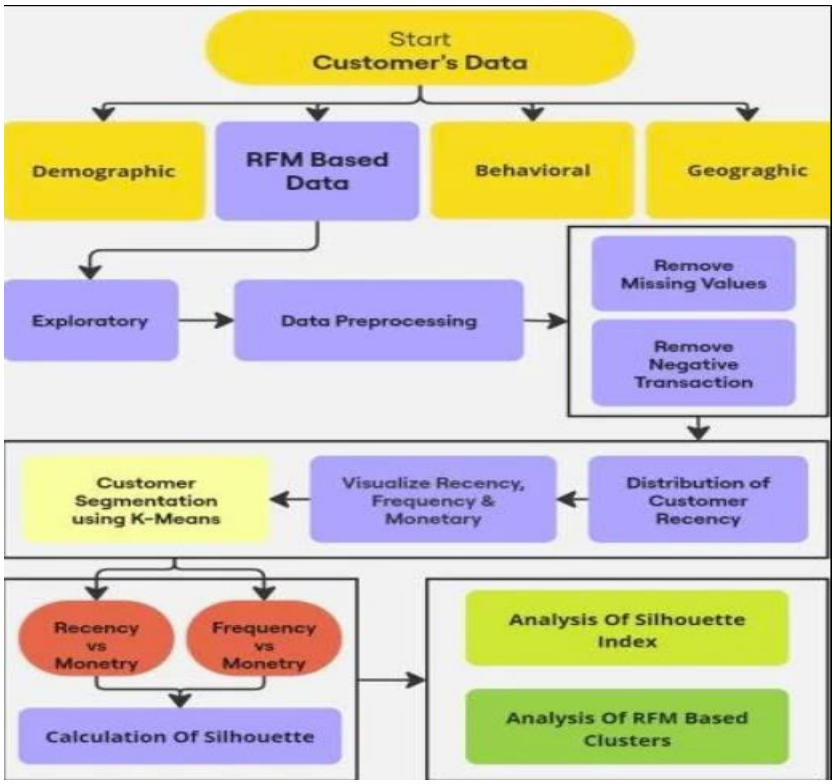
IV. Methodology

The methodology of the Customer Segmentation System involves a structured approach to analyze customer data and group customers into meaningful segments. Initially, a synthetic e-commerce dataset containing demographic, geographic, behavioral, and campaign-related features is collected. The dataset includes attributes such as age, income, purchase history, family size, and customer engagement.

The collected data undergoes preprocessing, which includes handling missing values, removing duplicates, encoding categorical variables, and normalizing numerical features using scaling techniques. After preprocessing, Exploratory Data Analysis (EDA) is performed to understand data distributions, correlations, and trends in customer behavior.

Feature engineering is applied to derive useful attributes such as total spending, purchase frequency, and customer activity levels.

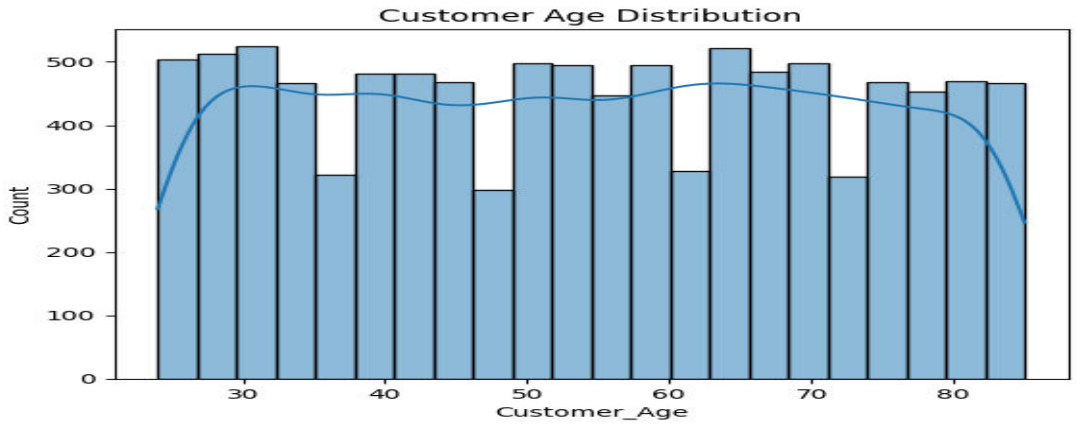
System Architecture

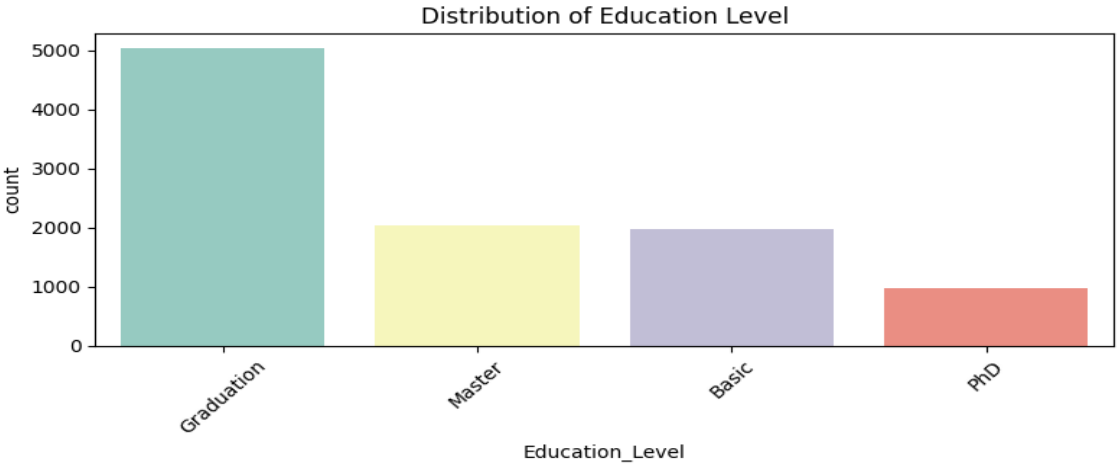


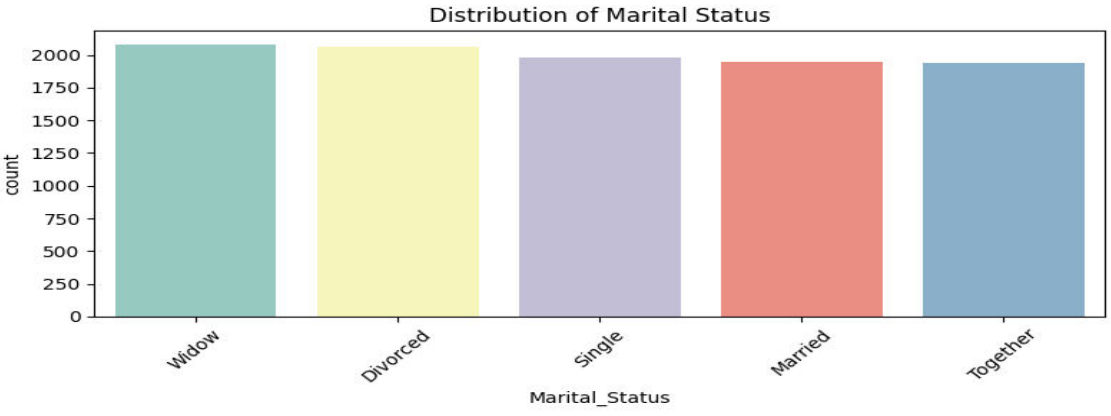
The system architecture defines the overall workflow of the Customer Segmentation System and how different components interact.

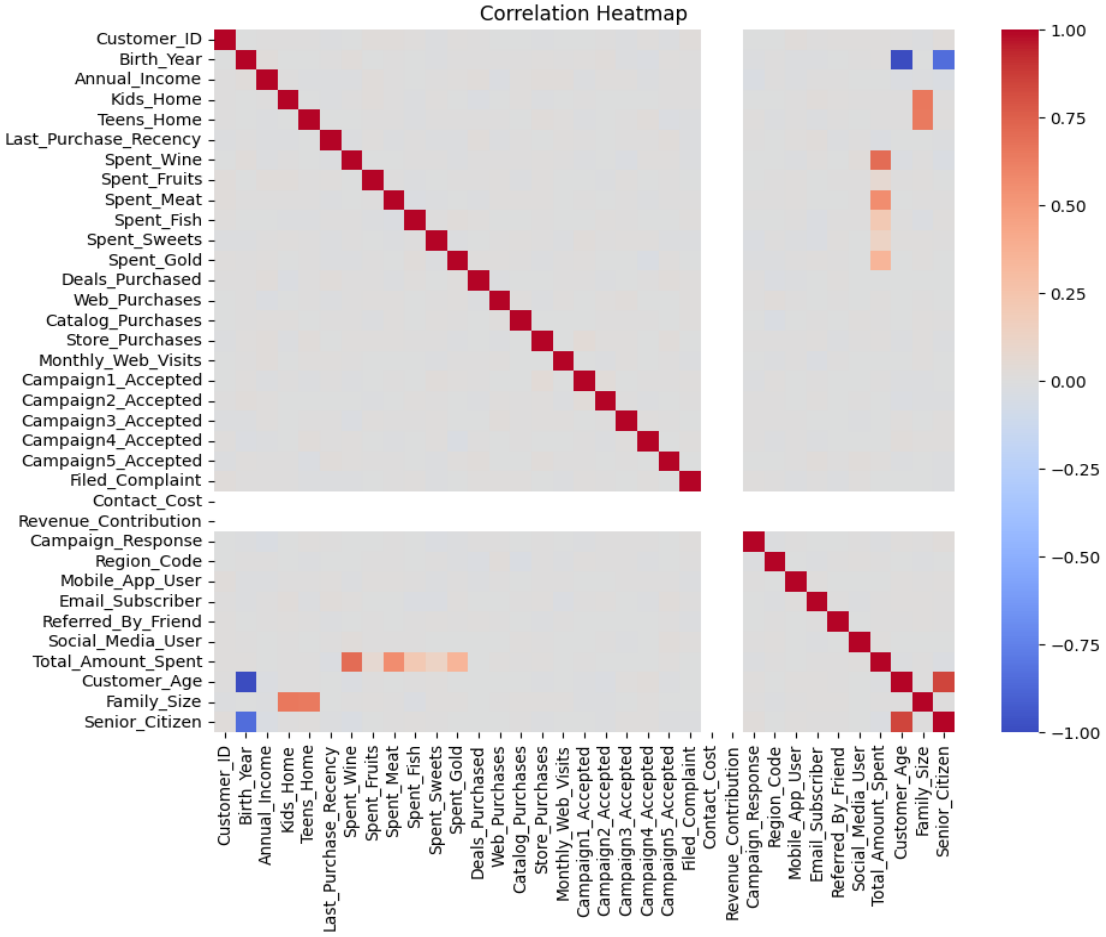
The process begins with the Data Source, which contains customer data such as demographic details, transaction history, and behavioral information. This data is passed to the Data Preprocessing Module, where cleaning, encoding, and feature scaling are performed.

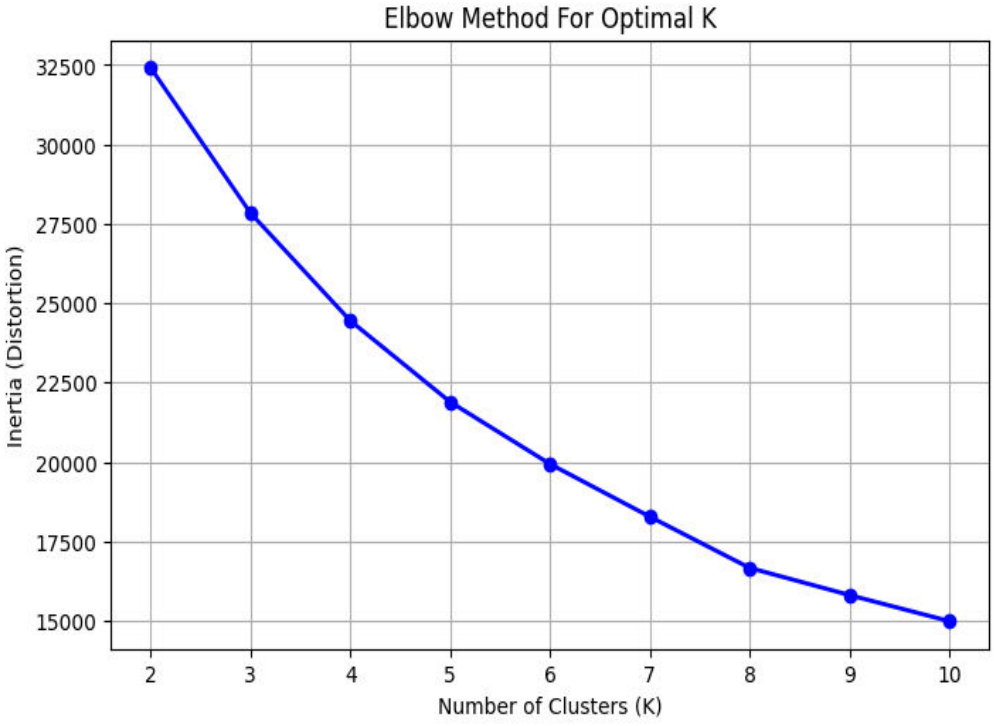
V. Result and Output

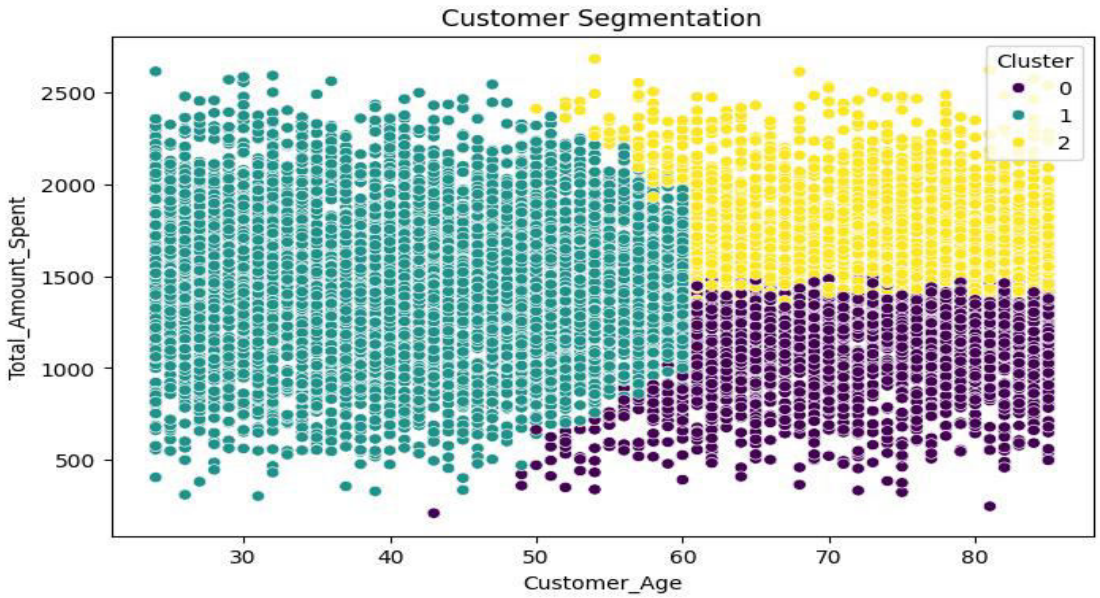







Correlation Heatmap
Customer_ID
Birth_Year
Annual_Income
Kids_Home
Teens_Home
Last_Purchase_Recency
Spent_Wine
Spent_Fruits
Spent_Meat
Spent_Fish
Spent_Sweets
Spent_Gold
Deals_Purchased
Web_Purchases
Catalog_Purchases
Store_Purchases
Monthly_Web_Visits
Campaign1_Accepted
Campaign2_Accepted
Campaign3_Accepted
Campaign4_Accepted
Campaign5_Accepted
Filed_Complaint
Contact_Cost
Revenue_Contribution
Campaign_Response
Region_Code
Mobile_App_User
Email_Subscriber
Referred_By_Friend
Social_Media_User
Total_Amount_Spent
Customer_Age
Family_Size
Senior_Citizen
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Elbow Method For Optimal K
Inertia (Distortion)
32500
30000
27500
25000
22500
20000
17500
15000
2
3
4
5
6
7
8
9
10
Number of Clusters (K)



VI. Conclusion

This study successfully implemented a customer segmentation model using the Scikit-learn K-Means clustering algorithm to group customers based on similar purchasing behavior. By analyzing customer data and identifying distinct segments, the model provides valuable insights into customer preferences, spending patterns, and engagement levels.

The use of visualization techniques such as Matplotlib and Seaborn enhanced the interpretability of the clusters, making it easier to understand and analyze customer groups. These insights enable businesses to design targeted marketing strategies, improve customer engagement, and optimize resource allocation.

The results demonstrate that unsupervised machine learning techniques like K-Means clustering are effective in extracting meaningful patterns from complex datasets. Overall, the project highlights the importance of data-driven decision-making in e-commerce and provides a practical approach for improving business performance through customer segmentation.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring
COVID-19 through intensive investigation with supervised machine learning algorithm. In
Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International
Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer
Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on
colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve
1Professor, Department of computer Science & engineering, Anurag University, TS, India.
2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time
Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT,
Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6,
doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis
of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial
Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications
in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer,
2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting
fake documents from land records using blockchain technology," in Blockchain for Smart
Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie
recommender using collaborative filtering and sentiment analysis," *International Research
Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025,
doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system
using LightGBM and decision tree machine learning models," *Journal of Machine and
Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, **Dr. B.V. Ramana Murthy**, "Smart agriculture through IoT and
machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and
Communication Engineering (ICCSCE)*, Apr. 2025.**[11]** Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer
learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199,
2024, doi: 10.1007/s42979-023-02500-x.